MMLSCU: A Dataset for Multi-modal Multi-domain Live Streaming Comment Understanding

Zixiang Meng
Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
Wuhan, China
zixiangmeng@whu.edu.cn

Qiang Gao
Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
Wuhan, China
gaoqiang@whu.edu.cn

Di Guo
Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
Wuhan, China
dylanguo@whu.edu.cn

Yunlong Li
Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
Wuhan, China
yunlongli@whu.edu.cn

Bobo Li
Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
Wuhan, China
boboli@whu.edu.cn

Hao Fei
National University of Singapore
Singapore, Singapore
haofei37@nus.edu.sg

Shengqiong Wu
Sea-NExT Joint Lab, National University of Singapore
Singapore, Singapore
swu@u.nus.edu

Fei Li*
Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
Wuhan, China
lifei_csnlp@whu.edu.cn

Chong Teng
Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
Wuhan, China
tengchong@whu.edu.cn

Donghong Ji
Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
Wuhan, China
dhji@whu.edu.cn

ABSTRACT

With the increasing popularity of live streaming, the interactions from viewers during a live streaming can provide more specific and constructive feedback for both the streamer and platform. In such scenario, the primary and most direct feedback method from the audience is through comments. Thus, mining these live streaming comments to unearth the intentions behind them and, in turn, aiding streamers to enhance their live streaming quality is significant for the well development of live streaming ecosystem. To this end, we introduce the MMLSCU dataset, containing 50,129 intention-annotated comments across multiple modalities (text, images, videos, audio) from eight streaming domains. Using multimodal pretrained large model and drawing inspiration from the Chain of Thoughts (CoT) concept, we implement an end-to-end model to sequentially perform the following tasks: viewer comment intent detection ⇒ intent cause mining ⇒ viewer comment explanation ⇒ streamer policy suggestion. We employ distinct branches for

*Fei Li is the corresponding author.

WWW '24, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0171-9/24/05...$15.00
https://doi.org/10.1145/3589334.3645677

video and audio to process their respective modalities. After obtaining the video and audio representations, we conduct a multimodal fusion with the comment. This integrated data is then fed into the large language model to perform inference across the four tasks following the CoT framework. Experimental results indicate that our model outperforms three multimodal classification baselines on comment intent detection and streamer policy suggestion, and one multimodal generation baselines on intent cause mining and viewer comment explanation. Compared to the models using only text, our multimodal setting yields superior outcomes. Moreover, incorporating CoT allows our model to enhance comment interpretation and more precise suggestions for the streamers. Our proposed dataset and model will bring new research attention on multimodal live streaming comment understanding.

CCS CONCEPTS

• Information systems → Multimedia information systems; **• Computing methodologies → Natural language generation**; **Language resources**.

KEYWORDS

Live Streaming, Multimodal, Comment Understanding

ACM Reference Format:
Zixiang Meng, Qiang Gao, Di Guo, Yunlong Li, Bobo Li, Hao Fei, Shengqiong Wu, Fei Li, Chong Teng, and Donghong Ji. 2024. MMLSCU: A Dataset for Multi-modal Multi-domain Live Streaming Comment Understanding. In *Proceedings of the ACM Web Conference 2024 (WWW '24), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3589334.3645677

1 INTRODUCTION

In the current era, live streaming has emerged as one of the dominant methods for content distribution, drawing a substantial number of streamers and viewers to participate. As depicted in Figure 1, which showcases the live streaming platform Twitch[1], streamers are delivering personalized live content to their viewers. Beyond merely watching, viewers actively post comments to engage in the live streaming sessions, expressing reactions for the live content. Such real-time comments also provide streamers valuable feedback, allowing for dynamic adjustments to content or strategies, thereby establishing a robust interaction loop between streamers and viewers. Given this context, mining the comments of live streaming holds practical value, contributing to both multimodal content understanding and the advancement of the live streaming industry.

However, it is difficult to understand and parse the unique community culture, as it encompasses a large amount of non-standard vocabulary, domain-specific jargon, memes, as well as oral expressions, and a variety of emojis [35]. In response to these challenges, some preliminary research has been conducted. Wang et al. [41] proposed a video comment multimodal dataset without any annotation, and the authors only suggested a comment generation task, lacking in-depth exploration of the content in the comment dataset. Similar issues exist in works such as [6, 28]. Xu et al. [45] introduced a live streaming dataset in the gaming domain, but the content of the single-domain community culture is limited, making it difficult to extend the model to other domains. Additionally, the authors determined audience preferences solely based on the number of comments, with limited research on the rich intentions contained within the comments. Similar single-domain live streaming dataset works can be found in [3, 4, 20]. The problem of single-domain focus and task specificity in these works hinders the study of the rich semantics embedded in live streaming comments, leading to a research gap in domain-independent and in-depth understanding of comment information.

To address the aforementioned issues, we constructed a multimodal, multi-domain live streaming comment dataset **MMLSCU** and conducted annotations on the comments. We proposed four tasks related to comment understanding:

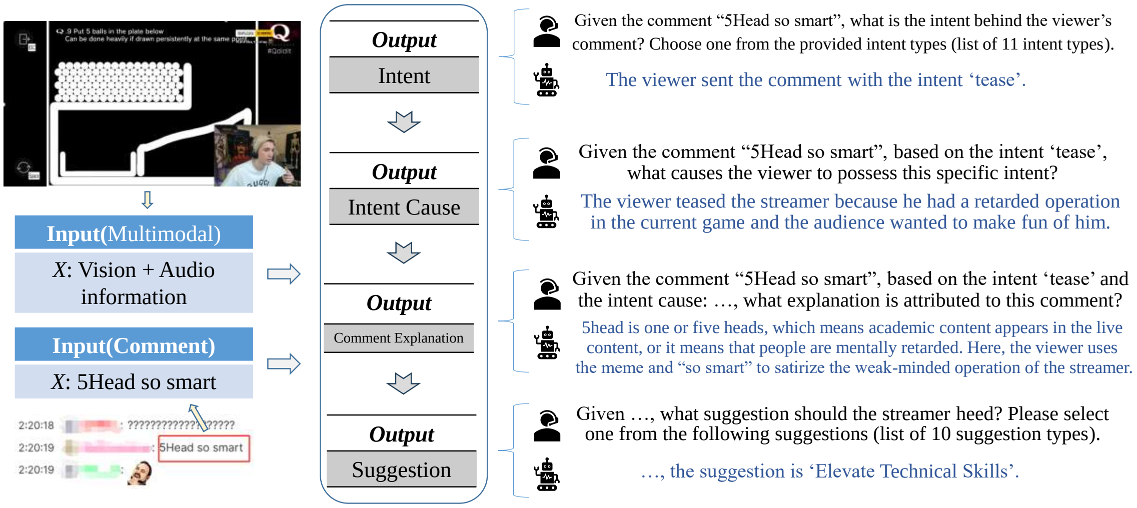
- Comment Intent Detection (CID): Discerning the underlying intent of comments and identifying hidden intentions for a deeper understanding of users' thoughts and needs.
- Intent Cause Mining (ICM): Seeking to ascertain the rationale behind a specific intent, analyzing the deeper psychological factors that drive users to express certain intentions.
- Viewer Comment Explanation (VCE): Generating in-depth explanations of comments from the viewer's perspective, and breaking down barriers imposed by specific community cultures.
- Streamer Policy Suggestion (SPS): Offering suggestions to streamers to help optimize content, adjust strategies, and increase user engagement.

Our dataset is sourced from Twitch live streams, encompassing video, audio, text, and emoji images. Our dataset comprises 8 domains, totaling 200 live segments, selected from 150 streamers. Streamer selection was balanced across factors such as age and region. Specific statistical information is provided in Table 5 of the Appendix. For the four tasks we proposed, corresponding annotations were conducted. For task 1 and 4, we designed 11 intent labels and 10 suggestion labels respectively, and performed multi-person cross-annotations. For task 2 and 3, we created intent cause and comment explanation texts. Through these four tasks, we aim to achieve a fine-grained analysis of real-time live streaming comments, uncover the hidden insights within live comments. This will provide enriched feedback to streamers, enhancing the quality of their streams and ultimately driving progress in the industry.

The four tasks present certain challenges. On one hand, it requires a unified approach to handle four different forms of subtasks, and there are interdependencies between these tasks. The performance of the preceding task affects the subsequent task. On the other hand, each task necessitates the integration of multimodal data over a period of time. For instance, due to cultural differences among different communities, only the video modality provides additional information to determine which culture a comment belongs to, and thereby infer the intent of the comment.

To address the previously mentioned challenges, we propose the Multi-Modal Four Comment Understanding Tasks (**MM^{4}CU**) model, which consists of three components: **(1) Video Branch**, which encompasses a pre-trained visual encoder designed to extract features from video frames, a position embedding layer to infuse temporal information, and a video Q-former for consolidating

[1] https://www.twitch.tv



Given the comment "5Head so smart", what is the intent behind the viewer's comment? Choose one from the provided intent types (list of 11 intent types).

The viewer sent the comment with the intent 'tease'.

Given the comment "5Head so smart", based on the intent 'tease', what causes the viewer to possess this specific intent?

The viewer teased the streamer because he had a retarded operation in the current game and the audience wanted to make fun of him.

Given the comment "5Head so smart", based on the intent 'tease' and the intent cause: …, what explanation is attributed to this comment?

5head is one or five heads, which means academic content appears in the live content, or it means that people are mentally retarded. Here, the viewer uses the meme and "so smart" to satirize the weak-minded operation of the streamer.

Given …, what suggestion should the streamer heed? Please select one from the following suggestions (list of 10 suggestion types).

…, the suggestion is 'Elevate Technical Skills'.

Figure 1: On the left, we give an example to show the live streaming scenario including a streamer, viewers and their comments. On the right, we give four examples to show the tasks, namely CID, ICM, VCE and SPS, in our proposed MMLSCU dataset.

frame-level representations. **(2) Audio Branch**, which involves a pre-trained audio encoder, a position embedding layer to incorporate temporal information into audio segments, and an audio Q-former for integrating diverse audio segment features. **(3) Text Decoder**, this component, for the fused multi-modal information, constructs the chain-of-thought (CoT) [42] for prompt learning. In our proposed model, we leverage CoT to tackle these tasks step by step, revealing the inherent relationships among them. Additionally, by harnessing the robust generative capabilities of large-scale models, we use Video-LLaMA [46] as the foundational model to effectively integrate features from various modalities.

We conducted a series of experiments on our MMLSCU dataset. Compared to the baseline, there were significant improvements in the F1 scores for both classification tasks: CID and SPS. Additionally, the metrics for the two generation tasks, ICM and VCE, also surpassed baseline by a large margin. Furthermore, our ablation experiments indicated that the introduction of multimodal information and CoT inference markedly enhanced the model's multi-step reasoning performance. The contributions of this paper include:

- We present a multi-modal, multi-domain live streaming comment dataset named MMLSCU. This dataset incorporates 4 distinct annotation types, facilitating the understanding of live streaming comments and furnishing streamers with nuanced feedback to elevate the overall streaming quality. To our knowledge, this work pioneers in filling this particular research void.
- Recognizing the associations of our proposed tasks, we architect a CoT framework to joint handle them and setup a strong benchmark for following-up work.
- Our analyses validate the superiority of utilizing multi-modal data compared to relying solely on text-based comment comprehension and feedback mechanisms.[2]

2 RELATED WORK

2.1 Live Streaming and Related Datasets

In recent years, live streaming has emerged as a pervasive phenomenon, particularly prominent within social media and the entertainment sector[37]. Empirical data suggests that most of the younger demographic has engaged with live streaming content at least once. Notably, several leading live streaming content creators have garnered a viewership that surpasses traditional television broadcasts[26]. This extensive viewership offers unparalleled opportunities for scholarly investigations into user engagement dynamics and intent recognition within the live stream-ing milieu. Furthermore, the evolving paradigms of digital gifting[21, 43] and bullet commentary, commonly referred to as "danmu"[15, 44], which are intrinsic to live streaming, present intriguing avenues for academic exploration.

Gaming-centric broadcasts are unequivocally recognized within the live streaming ecosystem as a predominant sub-domain. To facilitate a deeper understanding of user interactions within this context, a plethora of both unimodal[17, 18, 45] and multimodal datasets[3, 36, 38] have been curated. However, it is imperative to note that while gaming broadcasts occupy a pivotal position within the live streaming culture, they do not encapsulate its entirety. The spectrum of live streaming content is vast, encompassing domains such as casual conversations, educational sessions, culinary demonstrations, and travelogues, to name a few. These genres exhibit intrinsic disparities when juxtaposed with gaming broadcasts. Consequently, an exclusive reliance on datasets derived from gaming streams may not provide a holistic representation of the broader live streaming culture. While many unimodal datasets are tailored for the live streaming domain, notably those focusing on commentary text[1, 35], there is a discernible lacuna in the realm of comprehensive multimodal datasets. Several multimodal datasets are

[2] Our data and code are open at https://github.com/Newkic/MM4CU

tailored for short video segments and bullet commentary annotations[12, 41]. However, datasets that offer a comprehensive multimodal perspective on live streaming remain relatively sparse. Chen [6] proposes that MovieLC Dataset, a multimodal dataset tailored for the live streaming domain, is noteworthy. Yet, it predominantly aligns bullet commentaries with their corresponding video segments without delving into the underlying sentiment or the contextual triggers for such commentaries. Such nuanced information is pivotal for models aiming to better comprehend video content.

2.2 Multimodal Pretrained Models and CoT

The integration of textual and visual information has become a prominent research direction, leading to the emergence of several multimodal pretrained models. Building on the foundational success of unimodal architectures like BERT[7] for text and ResNet[14] for images, recent models aim to jointly learn representations across both modalities. Notably, CLIP[32] learns visual concepts from natural language supervision, demonstrating robust zero-shot performance across various visual benchmarks. Similarly, ViLBERT[25] employs a dual-stream approach, processing visual and textual inputs separately and then merging them, showcasing impressive results in visual question answering and visual commonsense reasoning.

Among these advancements, Large Language Model (LLM), such as ChatGPT[29], stand out for their approach to human-level intelligence[33]. VideoLLM[46] distinguishes itself by adeptly integrating visual and textual information from videos, emphasizing the narrative structure and temporal dynamics, proving its efficacy in tasks requiring a profound understanding of video content. Furthermore, there's compelling evidence that LLM possess an exceptional aptitude for common-sense understanding[24, 31].

Transitioning from their intrinsic understanding abilities, the introduction of CoT technique has gained prominence[47]. CoT has been widely used to enhance the multi-step reasoning capabilities of LLM by encouraging them to generate intermediate reasoning chains, guiding them towards problem solutions[42]. Notably, CoT prompting is a gradient-free approach that coaxes these models into articulating the intermediate steps leading to the final answer.

3 THE MMLSCU DATASET

The overall process of constructing the MMLSCU dataset is illustrated in Fig 2. In this section, we will introduce the details of data preparation, task definitions and data annotation, respectively.

3.1 Data Preparation

Due to the high quality, diversity, and wide viewer appeal of Twitch live streaming, we have selected Twitch as our data source for live streaming research. We enlisted the services of 14 seasoned viewers, well-versed in diverse online streaming scenarios, to observe live broadcasts or replays on Twitch for over a month. These scenarios include 8 domains: **Games**, **Just Chatting**, **In Real Life**, **Music and Performances**, **E-sports**, **Creative and Arts**, **Education and Learning**, and **Special Events**. In our dataset, 200 English live streaming clips since 2020 were selected, across total 2374 minutes and accompanied by 50,129 comments. The content of these clips is described and recorded at intervals of 20 seconds.

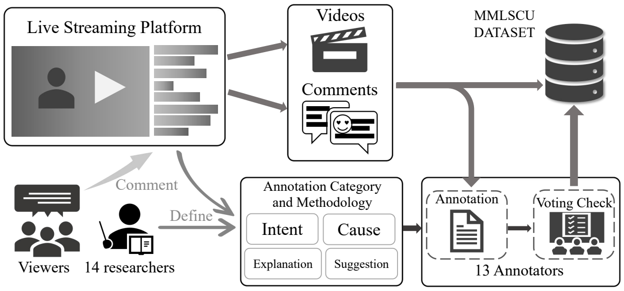


Figure 2: The overall process of constructing the MMLSCU dataset.

To ensure quality content and meanwhile keep diversity, the data selection strategy considers various factors such as streamer age, gender, and viewer comment count. Pertinent factors are stored in the meta-information corresponding to each live streaming clip in the dataset. To avoid phrases and emoticons from being heavily spammed in the streaming chat we collected, spam detection and cleaning was conducted.

Ethical Consideration We utilized the Twitch Developer API[3], to obtain live streaming clips, strictly adhering to fair use principles. As per Section 8 of Twitch's terms of service[4], viewer comments (live chat during streaming) are unlicensed, permitting us to process them. After data acquisition, any sensitive or unsavory comment information was removed. Furthermore, to protect privacy information, all data underwent de-identification, removing viewer identity information irrelevant to live streaming content.

3.2 Task Definitions

3.2.1 ***Comment Intent Detection (CID).*** We devised a novel multi-modal live streaming comment intent detection classification method, encompassing 4 coarse-grained and 11 fine-grained intent categories. For 4 coarse-grained label types, we determine types "Achieve Goals" and "Express Emotions" based on human intention philosophy[5], and the coarse-grained type "Platform Operational Interactions" based on Twitch's description for viewers.[5] The remaining comments are categorized as "Another" if they are empty or inconvenient to classify. We randomly sampled each domain, using coarse-grained intent types for preliminary labeling. We discovered that the coarse-grained intent labels were overly broad when describing intents in complex live-streaming scenarios. Consequently, we refined each coarse-grained type into finer granularities, pre-labeled the randomly sampled data, and merged labels with ambiguous distinctions, resulting in 11 fine-grained label types. The detailed explanations of these labels can be found in Appendix C.1.

3.2.2 ***Intent Cause Mining (ICM).*** Relying solely on intent labels is insufficient to comprehensively describe the motivations and psychological states behind the viewer's comments. Therefore, we proposed a novel task to analyze the intent cause behind a

[3] https://dev.twitch.tv/docs
[4] https://www.twitch.tv/p/en/legal/terms-of-service/
[5] https://www.twitch.tv/watch/

comment, revealing the reasons prompting viewers to make specific comments in the live streaming context, facilitating the comprehension of semantic significance conveyed in associated comments. Due to the non-explicit nature of intent behind comments in live streaming, "intent cause" most of the time cannot be extracted directly from comments, and mining "intent cause" requires consideration of the live streaming scenario relevant to the current comment. We utilize a generative approach to obtain the "intent cause", and regarding this task, we have manually written "intent cause" based on the comments and their relevant live streaming scenarios as the label data with the assistance of GPT4. The generated cause was subsequently filtered based on the criteria shown in the Appendix A.3.

3.2.3 ***Viewer Comment Explanation (VCE).*** Comment explanation is devised to delve deeply into the inherent meanings behind comments, interpreting not merely the detailed meaning of sentences but also amalgamating the prevailing live streaming context to interpret from the viewer's perspective holistically. Let the Comment Explanation CE be defined as $CE = SI + CI$, where SI represents the intrinsic meaning of a sentence considering only the text modality, and CI involves analyzing the information conveyed by the comment from the perspective of the viewer, incorporating multimodal information from the live streaming context. Since SI only considers the text modality, we consider using GPT-4 to assist in generating SI, followed by subsequent manual filtering. However, the part involving the use of multimodal information is in CI, for which we are not using GPT-4 for generation. The acquisition of CI is done through manual annotation. CI annotation steps are as follows: (1) describe the deep meaning of comments by integrating multimodal information ; (2) describing the message intended to be conveyed from the perspective of the viewer. The annotation example for Viewer Comment Explanation can be found in Appendix A.3.

3.2.4 ***Streamer Policy Suggestion (SPS).*** Upon a profound understanding and analysis of live streaming comments, furnishing precise suggestions to streamers is a pivotal step to harnessing viewer feedback judiciously. The 10 fine-grained suggestion labels were defined after referencing the Twitch Viewer Feedback Survey experiment[6] and Twitch Creator Camp[7]. The detailed explanations of these labels can be found in Appendix C.2. Most comments implicitly contain suggestions for the streamer. Comments that are without suggestions or have unclear suggestion types are labeled as **None**. We conducted two rounds of annotation. In the first round, we annotated fine-grained suggestion types for each comment. In the second round, we determined whether each suggestion is representative of the live streaming segment it belongs to. This was done by considering the first-round suggestion labels and related multimodal information. The annotator selects the segment that exhibits thematic consistency and annotates representative suggestions.

3.3 Data Annotation

Following the data preparation and annotation definitions, we engaged 13 personnel, all fervent and adept with the live streaming environment, to undertake the task of data annotation. Staff members were equipped with exemplary instances for each annotation type to serve as guiding benchmarks. Only those who underwent comprehensive training were permitted to annotate. To amplify the efficiency of the annotation process, we established a dedicated database to manage all multi-modal data and a user-friendly annotation interface. The team was bifurcated into two distinct units: an eight-member annotation team and a five-member review panel.

Our study encompasses four tasks, comprising two classification and two generation tasks. The annotation process is structured in four stages: first, annotate intent; then, annotate intent cause; next, annotate comment explanation; finally, annotate suggestion. The annotation is conducted while observing the live streaming video, meaning that the annotation results are derived from the multimodal information of the live streaming. Upon completion of the annotation process, a review team assessed the results through a voting mechanism. A label was deemed accepted if it garnered approval from three or more reviewers, achieving a three-fifths majority. Labels failing to achieve this majority were returned to the annotators for further refinement, pending majority approval in a subsequent round of review. When annotating emoticons, referencing external platforms, such as Know Your Meme[8], enhancing annotation precision. In Appendix A.1, we provide specific distribution details for two classification task labels and statistics regarding the number of comments in different domains.

4 METHODOLOGY

Utilizing a multi-modal form of large language models, we engineered a CoT framework expressly tailored for the tasks we had delineated.

4.1 Model Architecture

According to literature [45] focusing on observing live E-sports games on the Twitch platform, it becomes imperative to account for the time spectators invest in crafting their comments while watching the live streaming. This consideration arises due to inherent delays, such as typing time, which imply that comments posted by viewers at a given instance frequently pertain to live streaming content a few moments prior. In a user study documented by Palin[30], the average typing speed on keyboards, denoted as WPM_k, is found to be 52 words per minute, while the average typing speed on smartphones, referred to as WPM_s, is 38 words per minute. Incorporating both the WPM_k and WPM_s metrics to obtain the time of live streaming content related to the current comment. For any current comment, we obtain its time as T_c. Given that viewers cannot foresee forthcoming live scenarios, we consider the start time, T_s, of the segment of live streaming content corresponding to that particular comment. Define l as the prior duration of live streaming content associated with the current comment. l can be

[6] https://link.twitch.tv/LiveViewerFeedback

[7] https://www.twitch.tv/creatorcamp

[8] https://knowyourmeme.com/

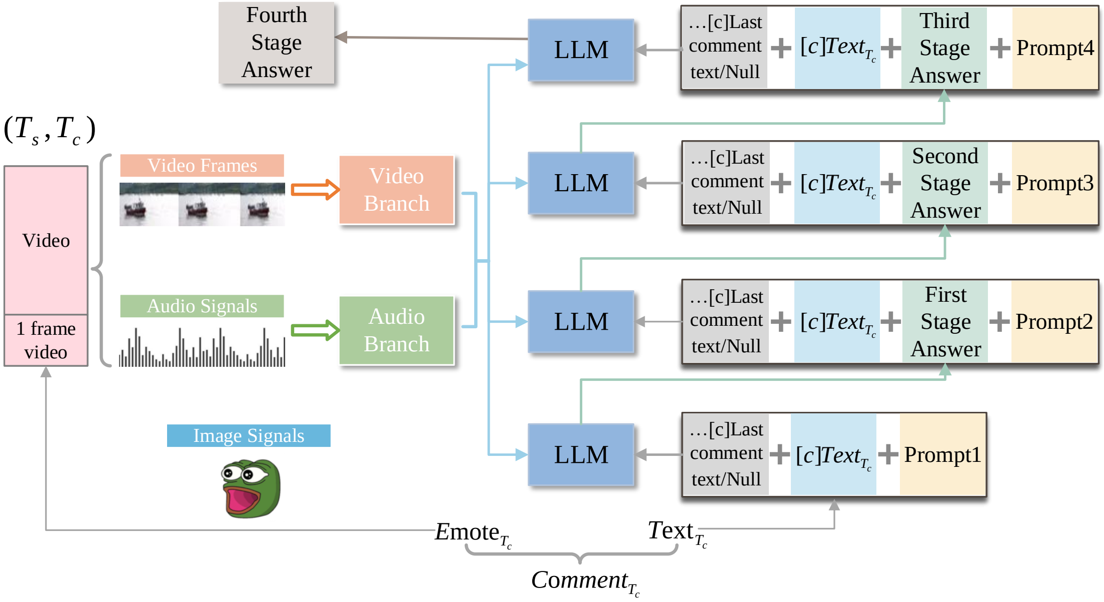


Figure 3: The architecture of our MM^{4}CU model for joint multimodal live streaming comment understanding with CoT schema.

determined using the given formula:

$$l = n_w \Big/ \frac{0.5 * (WPM_k + WPM_s)}{60}, \quad (1)$$

where n_w is the number of words in the current comment. The formula of T_s is as follows:

$$T_s = MAX(T_c - l, 0). \quad (2)$$

Therefore, considering a comment timestamped at T_c, the corresponding video and audio modalities should encompass live streaming content within the time frame (T_s, T_c). We interpret the image as a single video frame; thus, the emoticon $Emote_{T_c}$ in a comment $Comment_{T_c}$ at time T_c is treated as one video frame concatenated after the video frames within (T_s, T_c). Subsequently, the video and audio representations are relayed to the Large Language Model (LLM) to align with the text embedding dimension. The LLAMA-2-13B-chat[39] model was utilized as our LLM. The Vision branch processes the video and derives its representation, while the Audio branch is utilized for audio representation.

Vision Branch The objective of the Vision branch is to facilitate the comprehension of visual input by the LLM. This branch encompasses a pre-trained visual encoder designed to extract features from video frames, a position embedding layer to infuse temporal information within the frames, a video Q-former to aggregate frame-level representations, and a linear layer tasked with projecting these video outputs to a dimension congruent with the LLM's text embeddings. For the visual encoding process, we incorporate the pre-trained visual component of BLIP-2 [22] as the frozen visual encoder, which includes a ViT-G/14 [8] from EVA-CLIP and a pre-trained Q-former. The position embedding layer, video Q-former, and linear layer are initialized randomly and fine-tuned to effectively bridge the output of the frozen visual encoder with that of the frozen LLM.

Audio Branch The Audio branch is constructed to enable the LLM to interpret audio inputs. It comprises a pre-trained audio encoder, a position embedding layer to embed temporal information into audio segments, an audio Q-former to amalgamate different audio segment features and a final linear layer to map the audio representation to the embedding space of the LLM. We employ the pre-trained Imagebind [9] as our audio encoder. Analogous to the video Q-former, the audio Q-former instills temporal information by appending learnable position embeddings to audio segments. It subsequently generates fixed-length audio features by calculating interactions between position-encoded audio segments. The architecture of the audio Q-former mirrors that of the video Q-former, and, ultimately, a linear layer maps these audio features to the embedding domain of the LLM.

4.2 Chain-of-Thought Prompting

Considering the interrelations among the four tasks, we designed a multi-modal version of the CoT framework. This framework encompasses four sequential phases, specifically tailored to handle the tasks: viewer comment intent detection > intent cause mining > viewer comment explanation > streamer policy suggestion. While all four stages employ a consistent model architecture, variations are introduced in the input and the output.

For a comment $Comment_{T_c}$ at timestamp T_c, its corresponding live streaming content is during the(T_s, T_c) interval. $Comment_{T_c}$ consists of $Emote_{T_c}$ and $Text_{T_c}$, where $Emote_{T_c}$ can be empty.

First Stage In the first stage, our input is represented as

$$X^1 = \{x^1_{text}, x_{vision}, x_{audio}\}, \quad (3)$$

where:

$$x^1_{text} = x^1_{last} \circ Text_{T_c} \circ Prompt^1, \quad (4)$$

- x^1_{text} denotes the text part of input in first stage.

- x_{vision} denotes the vision part of input.
- x_{audio} denotes the audio part of input.
- x^1_{last} denotes the text part of the comments preceding the current comment in the first stage.
- $\circ$ denotes the *concatenate* operation.
- $Text_{T_c}$ denotes the text component of the current comment.
- $Prompt^1$ denotes the prompt context of the first stage.

When the current comment is inaugural, x^1_{last} defaults to $NULL$. Otherwise, comments are demarcated by the delimiter $[C]$. If appending the text from a prior comment exceeds the maximum input length, the over-extending portion of that comment's text is truncated. In the first stage, The prompt template is articulated as:

> *Template – Stage 1*
>
> C_1[Given the comment $Text_{T_c}$ and its accompanying multi-modal live streaming data], what is the intent behind the viewer's comment? Choose one from the provided intent types (list of 11 intent types).

Here C_1 signifies the prompting context for the first stage. This can be formally expressed as I=argmax$p(i|Text_{T_c})$, where I represents the output text denoting the comment's intent, also visualized as the *First Stage Answer* in the figure 3.

Second Stage In the second stage, our input is represented as

$$X^2 = \{x^2_{text}, x_{vision}, x_{audio}\}, \tag{5}$$

where:

$$x^2_{text} = x^2_{last} \circ Text_{T_c} \circ First\ Stage\ Answer \circ Prompt^2, \tag{6}$$

In the second stage, The prompt template is articulated as:

> *Template – Stage 2*
>
> $C_2[C_1, I]$. Based on the identified intent, what causes the viewer to possess this specific intent?.

C_2 acts as the prompting context for the second stage, concatenating C_1 and I. Mathematically, this is represented by C=argmax$p(c|Text_{T_c}, i)$, where C is the textual answer encapsulating potential causes for the intent, or as illustrated in the figure 3, *Second Stage Answer*.

Third Stage In the third stage, our input is represented as

$$X^3 = \{x^3_{text}, x_{vision}, x_{audio}\}, \tag{7}$$

where:

$$x^3_{text} = x^3_{last} \circ Text_{T_c} \circ Second\ Stage\ Answer \circ Prompt^3. \tag{8}$$

In the third stage, The prompt template is articulated as:

> *Template – Stage 3*
>
> $C_3[C_2, C]$. Grounded on the cause of intent, what explanation can be attributed to this comment?

C_3 is the prompt context for the third stage, concatenating C_2 and C. This is mathematically framed as E=argmax$p(e|Text_{T_c}, i, c)$, wherein E is the text capturing potential explanations of the comment. E is termed the *Third Stage Answer* in the figure 3.

Fourth Stage: In the fourth stage, our input is represented as

$$X^4 = \{x^4_{text}, x_{vision}, x_{audio}\}, \tag{9}$$

where:

$$x^4_{text} = x^4_{last} \circ Text_{T_c} \circ Third\ Stage\ Answer \circ Prompt^4. \tag{10}$$

In the Fourth stage, The prompt template is articulated as:

> *Template – Stage 4*
>
> $C_4[C_3, E]$. Based on the comment explanation, what suggestion should the streamer heed? Please select one from the following suggestions (list of 10 suggestion types).

C_4 operates as the prompt context for the fourth stage, concatenating C_3 and E. S=argmax$p(S|Text_{T_c}, i, c, e)$, wherein S denotes the resultant suggestion text or the *Fourth Stage Answer* illustrated in the figure 3.

4.3 Training Strategy

For the pre-training of the Vision branch, we employed our live-stream clips along with their associated descriptive metadata. we incorporated a video-to-text generative task, wherein a 20-second live-stream video and its corresponding description served as inputs to prompt the frozen LLM to generate an apt text description. The objective of this phase was to leverage live-stream data to imbue the video features with as much live-stream scenario knowledge as possible. Given the scarcity of audio-text data, directly training the Audio branch posed significant challenges. The aim of the learnable parameters within the audio-language branch was to align the output embedding of the frozen audio encoder with the LLM's embedding space. After pre-training, our model was fine-tuned using our annotated dataset. For more training details, see Appendix B.3.

5 EXPERIMENTS

In this research, we propose a multimodal dataset designed to provide a robust foundation for studies in the field of live streaming. Upon finalizing our dataset, we proceeded to segregate it into training, validation, and test sets in an 8:1:1 ratio. To evaluate the effectiveness of our dataset, we conducted a series of experiments and compared the results with existing models. This section will introduce our experimental setup and the analysis of the experiments.

Table 1: Test set results for the CID and SPS tasks.

	CID				SPS			
	Acc_{11}	P	R	F_1	Acc_{10}	P	R	F_1
MAG-BERT	66.31	65.83	63.46	64.62	60.15	60.65	55.46	57.94
MUIT	64.59	62.48	66.24	64.31	59.98	58.03	56.07	57.03
MISA	65.72	63.05	65.57	64.29	59.04	57.73	54.80	56.23
MM^{4}CU	73.00	75.23	71.59	73.36	71.06	71.32	69.48	70.39

Table 2: Test set results for the ICM and VCE tasks.

	ICM				VCE			
	B^3	B^4	RO	ME	B^3	B^4	RO	ME
UniVL	18.23	12.61	22.15	23.76	15.32	10.11	19.74	22.35
MM^{4}CU	33.10	27.15	37.98	34.05	31.21	26.12	35.93	30.23

Table 3: The ablation studies for the CID and SPS tasks.

	CID				SPS			
	Acc_{11}	P	R	F_1	Acc_{10}	P	R	F_1
MM^{4}CU	73.00	75.23	71.59	73.36	71.06	71.32	69.48	70.39
only text	71.89	73.98	70.14	72.01	69.17	69.82	67.54	68.66
- Vision	72.14	74.61	70.82	72.67	70.32	70.65	67.83	69.21
- Audio	72.61	74.83	71.11	72.92	70.66	70.90	68.75	69.81
- CoT	73.00	75.23	71.59	73.36	66.30	67.77	63.09	65.35

Table 4: The ablation studies for the ICM and VCE tasks.

	ICM				VCE			
	B^3	B^4	RO	ME	B^3	B^4	RO	ME
MM^{4}CU	33.10	27.15	37.98	34.05	31.21	26.12	35.93	30.23
only text	31.23	25.38	35.42	32.17	29.40	24.78	33.76	28.17
- Vision	32.01	26.12	36.91	33.09	29.97	25.06	34.50	29.11
- Audio	32.60	26.65	37.48	33.55	30.71	25.76	34.03	28.65
- CoT	30.89	23.40	33.65	31.04	26.96	23.22	31.34	27.79

5.1 Experimental Setup

Baseline To assess the performance of existing methods on our dataset, we conducted a series of experiments. For the intent detection and policy suggestion tasks, we selected the following models: **MAG-BERT** [34], **MulT** [40], and **MISA** [13]. For the intent cause mining and viewer comment explanation tasks, which are two generative tasks, we conducted experiments using the multimodal generative model **UniVL** [27].

Evaluation Metrics We used various evaluation metrics to assess the model performance. For the classification tasks, following the MulT [40] framework , we reported n-class accuracy (Acc_{11} for intent detection score classification, Acc_{10} for policy suggestion score classification), F_1 score, precision (P), and recall (R), calculated using macro-averaging[10]. For the generative tasks, we reported evaluation metrics such as B^3 and B^4, ROUGE-L (RO)[23], METEOR (ME)[2].

5.2 Main Result

The experimental results for the classification tasks are shown in Table 1 (all results are macro-averaged values). The results of the generation task are shown in Table 2.

From the experimental results, it is evident that our model has achieved a significant improvement compared to the baseline models. This improvement can be attributed to our innovative multimodal architecture and the powerful inferential capabilities of the large text model. Furthermore, it can be observed that the classification performance for the "streamer policy suggestion" task is lower than that of the "comment intent detection" task. This is because the streamer policy suggestion is our fourth task, and its results are influenced by the outcomes of the preceding three tasks. Additionally, making policy suggestions requires the synthesis of information from a previous time period, making it a more challenging task compared to intent classification, hence resulting in lower performance metrics.

5.3 Ablation Studies

In assessing the influence of different modal information on classification and generation tasks, we carried out additional ablation experiments. The results of these experiments are detailed in Table 3 and Table 4, for classification and generation tasks respectively. From the experimental results, it can be observed that removing either video or audio information results in a slight decrease in model performance. In the case where only the text modality is available, both classification tasks see a decrease of 1.11 and 1.89 in accuracy and a decrease of 1.35 and 1.73 in F1 score, respectively. The generative task metrics also show some decline, indicating that video and audio modalities indeed provide essential information for the classification tasks.

Additionally, as the intent classification task serves as the first step in our reasoning process, removing the CoT doesn't affect the model's performance. However, the subsequent three tasks rely on the CoT provided in the previous step, and removing the CoT results in a significant performance drop. This further underscores the importance of CoT in the multi-step reasoning process. Details of removing COT refer to Appendix B.1. We also conducted experiments to consolidate the four stages into a single stage, as referenced in Appendix B.2. The experimental results indicate that the performance of the single-stage approach is inferior to that of the four-stage approach across all four tasks.

6 CONCLUSION

Our research has created MMLSCU, a multimodal and cross-domain live streaming comment dataset, along with four comment understanding tasks. These tasks include Comment Intent Detection, Intent Cause Mining, Viewer Comment Explanation, and Streamer Policy Suggestion. Through experimentation, we have demonstrated that the introduction of multimodal data and CoT reasoning significantly improves model performance. This research fills a gap in domain-independent and in-depth comment information understanding, providing essential tools for enhancing live streaming quality and driving industry development. We will openly share our dataset and code to encourage more researchers to participate in future studies and further advance this field.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (No. 2022YFB3103602), the National Natural Science Foundation of China (No. 62176187), the open project of Sichuan Provincial Key Laboratory of Philosophy, the Social Science for Language Intelligence in Special Education (No. YYZN-2023-1) and CCF-Baidu Open Fund.

REFERENCES

[1] Emna Baccour, Aiman Erbad, Kashif Bilal, Amr Mohamed, Mohsen Guizani, and Mounir Hamdi. 2020. FacebookVideoLive18: A Live Video Streaming Dataset for Streams Metadata and Online Viewers Locations. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*. 476–483. https://doi.org/10.1109/ICIoT48696.2020.9089607

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[3] Anna Belova, Wen He, and Ziyi Zhong. 2019. E-Sports Talent Scouting Based on Multimodal Twitch Stream Data. *CoRR* abs/1907.01615 (2019). arXiv:1907.01615 http://arxiv.org/abs/1907.01615

[4] Florian Block, Victoria Hodge, Stephen Hobson, Nick Sephton, Sam Devlin, Marian F Ursu, Anders Drachen, and Peter I Cowling. 2018. Narrative bytes: Data-driven content production in esports. In *Proceedings of the 2018 ACM international conference on interactive experiences for TV and online video*. 29–41.

[5] Michael Bratman. 1987. Intention, plans, and practical reason. (1987).

[6] Jieting Chen, Junkai Ding, Wenping Chen, and Qin Jin. 2023. Knowledge Enhanced Model for Live Video Comment Generation. arXiv:2304.14657 [cs.CV]

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2022. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. arXiv:2211.07636 [cs.CV]

[9] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One Embedding Space To Bind Them All. arXiv:2305.05665 [cs.CV]

[10] Thamme Gowda, Weiqiu You, Constantine Lignos, and Jonathan May. 2021. Macro-average: rare types are important too. *arXiv preprint arXiv:2104.05700* (2021).

[11] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677* (2017).

[12] Vikram Gupta, Trisha Mittal, Puneet Mathur, Vaibhav Mishra, Mayank Maheshwari, Aniket Bera, Debdoot Mukherjee, and Dinesh Manocha. 2022. 3MASSIV: multilingual, multimodal and multi-aspect dataset of social media short videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21064–21075.

[13] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. *Proceedings of the 28th ACM International Conference on Multimedia* (2020). https://api.semanticscholar.org/CorpusID:218538102

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[15] Ming He, Yong Ge, Enhong Chen, Qi Liu, and Xuesong Wang. 2017. Exploring the emerging type of comment for online videos: Danmu. *ACM Transactions on the Web (TWEB)* 12, 1 (2017), 1–33.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[17] Hong Huang, Junjie H Xu, Xiaoling Ling, and Pujana Paliyawan. 2022. Sentence Punctuation for Collaborative Commentary Generation in Esports Live-Streaming. In *2022 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 1–2.

[18] Tatsuya Ishigaki, Goran Topić, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2021. Generating Racing Game Commentary from Vision, Language, and Structured Data. In *Proceedings of the 14th International Conference on Natural Language Generation*. 103–113.

[19] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[20] Athanasios Vasileios Kokkinakis, Simon Demediuk, Isabelle Nölle, Oluseyi Olarewaju, Sagarika Patra, Justus Robertson, Peter York, Alan Pedrassoli Pedrassoli Chitayat, Alistair Coates, Daniel Slawson, et al. 2020. Dax: Data-driven audience experiences in esports. In *ACM International Conference on Interactive Media Experiences*. 94–105.

[21] Yi-Chieh Lee, Chi-Hsien Yen, Dennis Wang, and Wai-Tat Fu. 2019. Understanding how digital gifting influences social interaction on live streams. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–10.

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597 [cs.CV]

[23] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[24] Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2021. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387* (2021).

[25] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[26] Zhicong Lu, Haijun Xia, Seongkook Heo, and Daniel Wigdor. 2018. You watch, you give, and you engage: a study of live streaming practices in China. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[27] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation. arXiv:2002.06353 [cs.CV]

[28] Shuming Ma, Lei Cui, Damai Dai, Furu Wei, and Xu Sun. 2018. LiveBot: Generating Live Video Comments Based on Visual and Textual Contexts. arXiv:1809.04938 [cs.CL]

[29] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[30] Kseniia Palin, Anna Maria Feit, Sunjun Kim, Per Ola Kristensson, and Antti Oulasvirta. 2019. How do people type on mobile devices? Observations from a study with 37,000 volunteers. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–12.

[31] Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Prompting contrastive explanations for commonsense reasoning tasks. *arXiv preprint arXiv:2106.06823* (2021).

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[34] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2359–2369. https://doi.org/10.18653/v1/2020.acl-main.214

[35] Charles Ringer, Mihalis A. Nicolaou, and James Alfred Walker. 2020. TwitchChat: A Dataset for Exploring Livestream Chat. In *Proceedings of the Sixteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE'20)*. AAAI Press, Article 37, 7 pages.

[36] Anton Smerdov, Bo Zhou, Paul Lukowicz, and Andrey Somov. 2020. Collection and validation of psychophysiological data from professional and amateur players: A multimodal esports dataset. *arXiv preprint arXiv:2011.00958* (2020).

[37] Thomas Smith, Marianna Obrist, and Peter Wright. 2013. Live-streaming changes the (video) game. In *Proceedings of the 11th european conference on Interactive TV and video*. 131–138.

[38] Tsunehiko Tanaka and Edgar Simo-Serra. 2021. Lol-v2t: Large-scale esports video description dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4557–4566.

[39] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[40] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6558–6569. https://doi.org/10.18653/v1/P19-1656

[41] Weiying Wang, Jieting Chen, and Qin Jin. 2020. VideoIC: A Video Interactive Comments Dataset and Multimodal Multitask Learning for Comments Generation. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 2599–2607. https://doi.org/10.1145/3394171.3413890

[42] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL]

Table 5: Statistics of the MMLSCU dataset.

Meta Item	Information
Total number of live streaming clips	200
Total number of comments	50,129
Total number of words in comments	183,755
Total number of emoticons in comments	30,712
Maximum number of words in comments	65
Average duration of live streaming clips(s)	712.38
Maximum duration of live streaming clips(s)	840.00
Average string length of comment texts	21.70
Maximum string length of comment texts	500
Total number of fields	8
Total number of streamers	150

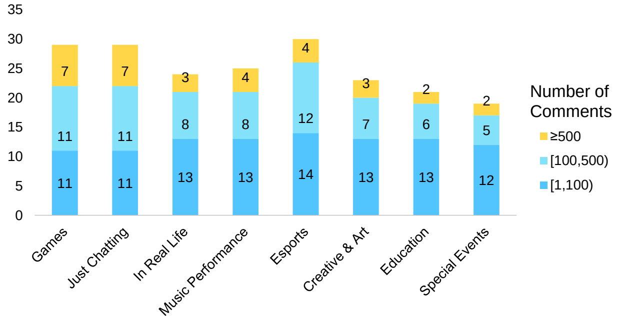


Figure 4: Distribution of domain and number of comments in live slices.

[43] Dinghao Xi, Liumin Tang, Runyu Chen, and Wei Xu. 2023. A multimodal time-series method for gifting prediction in live streaming platforms. *Information Processing & Management* 60, 3 (2023), 103254.

[44] Dinghao Xi, Wei Xu, Runyu Chen, Yuhang Zhou, and Zhan Yang. 2021. Sending or not? A multimodal framework for Danmaku comment prediction. *Information Processing & Management* 58, 6 (2021), 102687.

[45] Junjie H. Xu, Yu Nakano, Lingrong Kong, and Kojiro Iizuka. 2023. CS-Lol: A Dataset of Viewer Comment with Scene in E-Sports Live-Streaming. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (Austin, TX, USA) *(CHIIR '23)*. Association for Computing Machinery, New York, NY, USA, 422–426. https://doi.org/10.1145/3576840.3578334

[46] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. arXiv:2306.02858 [cs.CL]

[47] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923* (2023).

A DATASET DETAILS

A.1 Statistical Information

To further explore the dataset's domain characteristics and the distribution of labels for the two classification tasks, in Figure 4, we've provided a breakdown of comment counts across various live streaming domains. In Figure 5, we've presented the vote counts for different intent labels during the annotation process, and in Figure 6, we've shown the vote counts for various recommendation labels.

A.2 Prompt Template for Labeling Task

For intent cause mining and viewer comment explanation, we created the following prompt templates to generate labels.

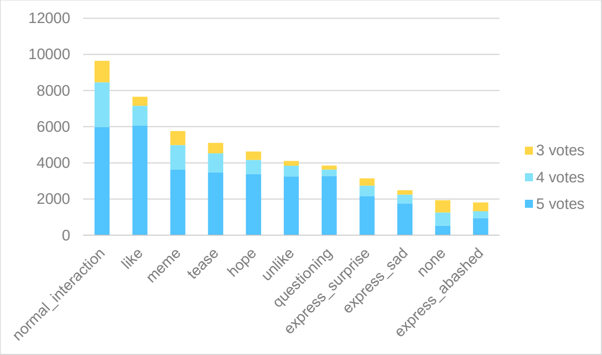


Figure 5: Intent type distribution and voting results

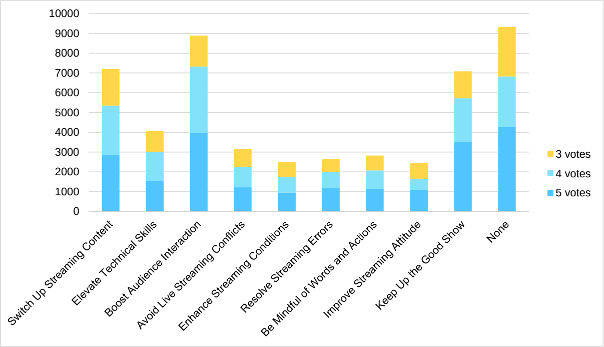


Figure 6: Suggestion type distribution and voting results

Template

Given a live streaming comment with the content [*comment*] and an associated intent of [*intent*], what is the cause behind this intent?

Template

Given a live streaming comment with the content [*comment*] , what is the explanation of the intrinsic meaning of the comment?

A.3 GPT4 Filtering Criteria and Generation Cases

GPT4 Filtering Criteria:

Relevance - the generated cause must exhibit a strong correlation with both the comment's content and its intent; Logical Consistency - the cause should align logically with the live streaming context and content; and Conciseness - the described cause should be succinct, eschewing undue complexity.

Wrong ICM Generation Cases:

Comment: *This music takes me back!*

Intent: Express_sad

Task: Intent Cause Mining

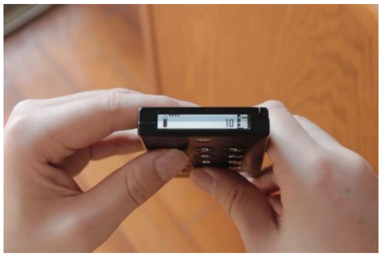


Figure 7: The streamer is showcasing a newly purchased electronic products.

GPT4's Response: *The user might be reminded of a popular dance trend from a few years ago due to the music.*

Error analysis: In a live streaming scene in the domain of music, the streamer sang a very touching old song that could make the viewer cry, evoking the viewer's sad feelings of missing the past and feeling that time has passed and things have changed. The comments of the live streaming were all expressing such sadness, but the answers generated by GPT4 were obtained only by comments.

Acceptable SI Generation Case:

Comment: *That technique is straight out of the 90s!*

Task: Viewer Comment Explanation.

SI: *The technique being demonstrated or discussed is reminiscent of outdated technology or things from the 1990s.*

Live streaming scenario: As shown in the figure 7

CI: *The viewer believes that the product appears dated and similar to older technique from the 1990s. This hints at a criticism that this product is not as advanced as streamer might think and is not worth it.*

CE: *The technique being demonstrated or discussed is reminiscent of outdated technology or things from the 1990s. The commenter believes that the product appears dated and similar to older technique from the 1990s. This hints at a criticism that this product is not as advanced as streamer might think and is not worth it.*

B FURTHER EXPERIMENTAL ANALYSIS

B.1 CoT Ablation Studies

Details of removing COT are as follows: We retain only the first phase's prompting context C_1 regarding input comments and multimodal information, and remove all intermediate results I, C, E, and guiding information C_2, C_3, C_4 from subsequent stages. We directly requested the model to provide results, as illustrated in the following example:

Stage 3 before removing COT: $C_3[C_2, C]$. Grounded on the cause of intent, what explanation can be attributed to this comment?

Stage 3 after removing COT： C_1[Given the comment and its accompanying multi-modal live streaming data], what explanation can be attributed to this comment?

B.2 Four Stages To Single Stage Experiment

To consolidate the four stages into a single-stage input question, we conducted additional experiments and found that the single-stage approach indeed performs worse than the multi-stage one. The specific experimental results are show in Table 6, Table 7 and Table 8. According to the experimental results, it can be observed that the performance of the single-stage approach is inferior to that of the fourth-stage approach across all four tasks. This indicates that the progressive inference in the four-stage process indeed guides the model's reasoning path.

Table 6: The single-stage experiment for the CID and SPS tasks on the MMLSCU test set.

	CID				SPS			
	Acc_{11}	P	R	F_1	Acc_{10}	P	R	F_1
Four-stage	73.00	75.23	71.59	73.36	71.06	71.32	69.48	70.39
Sing-stage	72.78	75.01	70.15	72.50	67.48	69.12	65.11	67.06

Table 7: The single-stage experiment for the ICM task on the MMLSCU test set.

ICM	B^3	B^4	RO	ME
Four-stage	33.10	27.15	37.98	34.05
Single-stage	31.02	23.38	33.97	32.13

Table 8: The single-stage experiment for the VCE task on the MMLSCU test set.

VCE	B^3	B^4	RO	ME
Four-stage	31.21	26.12	35.93	30.23
Single-stage	27.84	24.09	32.00	28.12

B.3 Training Details

Training on the annotated dataset began with pre-training on the Vision Branch and Audio Branch, utilizing collected live streaming videos and their corresponding descriptions. During this phase, the weights of Llama2 were frozen. The Vision Branch and Audio Branch received training to enhance understanding of live streaming videos and audio in the domain. Following pre-training, the model underwent fine-tuning with the annotated data. At this stage, the weights of the other branches were frozen, and LoRA[16] fine-tuning was performed on the LLM. We trained our model on 2 NVIDIA Tesla A100 GPUs. A learning rate warm-up strategy[11] was employed, starting from an initial learning rate of 0.0001 and linearly increasing to 0.001, beyond which it remained constant. The Adam optimizer[19], with β1 set to 0.9 and β2 set to 0.999, was used for optimization. The training process spanned a total of 10 epochs.

C LABEL EXPLANATION

C.1 Label Explanation for CID Task

The labels explanation for CID task are in Table 9.

C.2 Label Explanation for SPS Task

The labels explanation for SPS task are in Table 10.

Table 9: The explanation for CID task’s labels.

	Intent	Explanation
Achieve Goals	like unlike hope questioning tease	Like, Support, Enjoy, Comfortable, Pleasant Dislike, Oppose, Threaten, Irrational Hope, Suggestion, Spectator Question, Doubt, Confusion Mock, Ridicule
Express Emotions	express_surprise express_sad express_abashed	Expressing surprise or astonishment Expressing sadness or regret Expressing awkwardness or embarrassment
Platform Operational Interactons	normal_interactoin	The normal interaction in a livestream room
	meme	Recreational or straightforward meme play in the livestream room
Another	none	No comment posted or unclear intent

Table 10: The explanation for SPS task’s labels.

	Suggestion	Explanation
Content Strategy	Switch Up Streaming Content	The viewer finds the current livestream content too dull and suggests switching to or trying out new content.
	Elevate Technical Skills	The viewer thinks the streamer is not skilled enough and suggests that the streamer should improve their technical proficiency.
Engagement Strategy	Boost Audience Interaction	The viewer feels that the streamer lacks interaction with them, verlooks their opinions and requests, and suggests that the streamer should enhance interaction with the viewer.
	Avoid Live Streaming Conflicts	The streamer or certain viewer members in this livestream room are engaging in provocative or conflict-inducing behavior. It is advised that the streamer takes steps to avoid conflicts during the livestream.
Streaming Environment	Enhance Streaming Conditions	The viewer feels that the streamer’s livestream equipment conditions are subpar, or there are issues with background noise. They suggest that the streamer should improve the livestreaming environment.
	Resolve Streaming Errors	The streamer is currently experiencing issues with the livestream, such as network problems or a disabled camera.It is suggested that the streamer promptly address and resolve these errors.
Streaming Ethics	Be Mindful of Words and Actions	The viewer is warning the streamer about discussing or engaging in inappropriate topics or actions, such as skirting the edges or promoting racial discrimination. They emphasize the importance of the streamer being mindful of their words and actions.
	Improve Streaming Attitude	The viewer feels that the streamer is not putting enough effort into the livestream and seems distracted. They suggest that the streamer should livestream with more dedication and a focused mindset, and correct their attitude.
	Keep Up the Good Show	The viewer is very satisfied with the current program and encourages the streamer to keep up the good work. They hope the streamer will continue to maintain this level of performance.
	None	